A Sampling Approach to Generating Closely Interacting 3D Pose-pairs from 2D Annotations

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2018.2832097, IEEE Transactions on Visualization and Computer Graphics

A Sampling Approach to Generating Closely Interacting 3D Pose-pairs from 2D Annotations

Kangxue Yin, Hui Huang*, *Member, IEEE,* Edmond S. L. Ho,
Hao Wang, Taku Komura, Daniel Cohen-Or, Hao Zhang, *Senior Member, IEEE*

Abstract—We introduce a data-driven method to generate a large number of *plausible*, *closely interacting* 3D human pose-pairs, for a given motion category, e.g., wrestling or salsa dance. With much difficulty in acquiring close interactions using 3D sensors, our approach utilizes abundant existing video data which cover many human activities. Instead of treating the data generation problem as one of reconstruction, either through 3D acquisition or direct 2D-to-3D data lifting from video annotations, we present a solution based on Markov Chain Monte Carlo (MCMC) sampling. Given a motion category and a set of video frames depicting the motion with the 2D pose-pair in each frame annotated, we start the sampling with one or few seed 3D pose-pairs which are manually created based on the target motion category. The initial set is then augmented by MCMC sampling around the seeds, via the Metropolis-Hastings algorithm and guided by a probability density function (PDF) that is defined by two terms to bias the sampling towards 3D pose-pairs that are physically valid and plausible for the motion category. With a focus on efficient sampling over the space of close interactions, rather than pose spaces, we develop a novel representation called *interaction coordinates* (IC) to encode both poses and their interactions in an integrated manner. Plausibility of a 3D pose-pair is then defined based on the IC and with respect to the annotated 2D pose-pairs from video. We show that our sampling-based approach is able to efficiently synthesize a large volume of plausible, closely interacting 3D pose-pairs which provide a good coverage of the input 2D pose-pairs.

Index Terms—Closely interacting 3D human poses, data generation and augmentation, MCMC sampling.

1 Introduction

Studies of human pose geometries and motions are ubiquitous in geometry processing and analysis. The inherent biomechanical and neurophysiological complexities underlying human movements and behaviors are not always easy to comprehend through purely geometric analysis; this has motivated the development of many data-driven solutions to model and analyze human poses. Human pose or motion data in high volumes and large varieties have offered much benefit to numerous applications including motion synthesis [1], pose estimation [2], biped control [3], and geometry-based analyses of object functionalities [4], [5], [6].

There has been great progress on acquisition of 3D pose geometries for *single* or isolated characters, e.g., the CMU motion capture (MoCap) database, and some of the most successful reconstruction methods have been data-driven [2], [7]. However, the amount of data and related work which can capture *close interactions* between 3D human poses has been conspicuously small. With a greater degree of occlusion arising from close interactions, the ability of optical markers or view-based sensors (e.g., MS Kinect or video) to acquire quality data is severely impaired. Typically, state-of-the-art methods can only handle light interactions under slow motion. Inertial or magnetic MoCap may be viable options to remedy the occlusion problem. However, non-optical systems usually suffer from problems such as drifting, distortion and low precision. Moreover, as the sensors are usually placed on bulky body suits, the movement freedom of the performers is indeed highly compromised.

Figure 1. Closely interacting 3D wrestling poses automatically generated by MCMC sampling from a *single* seed pose (center).

In this paper, we introduce a method to generate a large volume of plausible, closely interacting 3D human pose-pairs, for a given motion category, e.g., classical wrestling; see Figure 1. With the great difficulties in acquiring close interactions in the 3D setting, one possible solution is to utilize abundant existing *video* which cover many human activities. However, even with sufficient image annotation over the video, lifting flat poses into 3D figures is already quite involved for single or isolated characters [8], [9], [10]. Most of these works rely on pose priors learned from a 3D MoCap dataset. Applying similar lifting schemes for interacting

- *K. Yin and H. Zhang are with Simon Fraser University*
- *H. Huang and H. Wang are with Shenzhen University*
- *E. Ho is with Northumbria University*
- *T. Komura is with University of Edinburgh*
- *D. Cohen-Or is with Shenzhen University and Tel Aviv University*

**Corresponding author: Hui Huang (hhzhiyan@gmail.com)*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2018.2832097, IEEE Transactions on Visualization and Computer Graphics

pose-pairs is more difficult, since occlusions and other issues with lifting and tracking are greatly amplified by close interactions. In addition, the 3D MoCap dataset from which these methods learn their pose priors may not even have certain poses that can only exist for interacting people, e.g., wrestling poses. To solve such an ill-posed problem, a large volume of 3D pose interaction data as a prior would have been valuable, but such data, which we are going after in this work, do not yet exist.

Instead of treating the data generation problem as one of reconstruction, either through 3D acquisition or direct 2D-to-3D data lifting, we make the key observation that we can solve the problem through a *sampling* process. If the generated sample 3D pose-pairs can reproduce input 2D frames, when projected along appropriate views, then we essentially achieve an indirect *lifting by sampling*; see Figure 2. Our primary sampling criterion is to ensure that the samples are plausible, as dictated by the input data for the motion category. However, owing to the stochasticity of the sampling process, we may obtain samples whose projections deviate sufficiently from all the input frames, offering the potential to generate *novel* interacting poses. Importantly, all of these can be accomplished without resorting to 3D MoCap databases.

Given a motion category and a set of video frames depicting the motion with the 2D *skeletal* pose-pair in each frame annotated, we start the sampling with one or more manually designed *seed* 3D skeletal pose-pairs, each of which is encoded by a vector of skeletal joint positions. The seed set is augmented via a Markov Chain Monte Carlo (MCMC) sampling over the space of 3D skeletal pose-pairs, around the seeds and guided by a probability density function (PDF) with two terms:

- A *physical prior* biases the sampling towards producing physically plausible individual poses in the 3D pose-pair.
- A *data-driven* plausibility or likelihood measure for a 3D pose-pair with respect to the 2D pose-pair data from video. Specifically, we encode the interaction between a pose-pair in 2D using a series of vectors among skeletal joints of the two poses, which we call *interaction coordinates* or IC. Then the plausibility measure for a 3D pose-pair is estimated, via the IC, by how close its various projections are to the 2D pose-pair for the motion category considered.

Based on the PDF, the sampling follows the well-established Metropolis-Hastings scheme. However, to improve sampling quality and further leverage the annotated video data, we introduce a *velocity bias* to steer Metropolis-Hastings. The velocity bias is applied to 3D pose-pairs, but it is based on velocity information extracted from the motion of 2D pose-pairs in the video.

We show that with only a single seed 3D pose-pair, our data augmentation scheme is able to synthesize a large volume of plausible, closely interacting poses through MCMC sampling for various motion categories. Importantly, we show that the sampling produces 3D pose-pairs with a good coverage of the input 2D data (via projection) efficiently. The coverage applies to both annotated 2D pose-pairs and un-annotated in-betweens.

Furthermore, with lifting by sampling, there is no need to directly lift any flat pose-pair to 3D, we only need to assess the physical validity and interaction plausibility of the (projected) 3D pose-pairs resulting from MCMC sampling. While direct 2D-to-3D pose lifting methods require a large 3D MoCap dataset to train their data-driven models, our method employs only 2D annotations over video which is much easier to obtain than 3D MoCap or annotated data in 3D. In Figure 2, we show a comparison of our lifting-by-sampling scheme to state-of-the-art methods [10], [11] for direct 2D-to-3D lifting of individual poses. To contrast the quality and plausibility of the generated interactions, i.e., how the hands and arms of the wrestlers are posed and making contacts with each other, we highlighted some regions in circles.

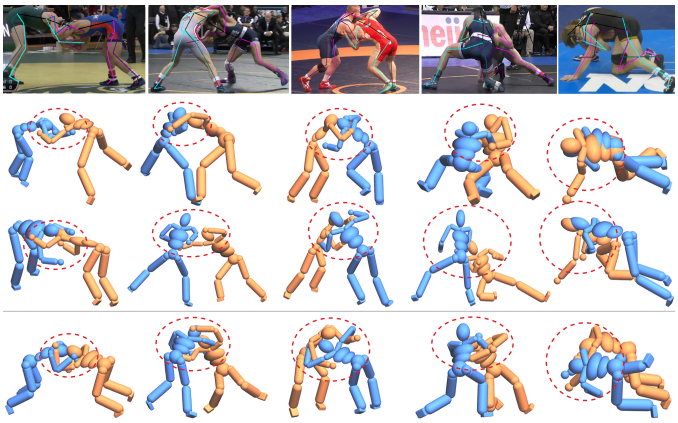


Figure 2. Direct 2D-to-3D lifting (second and third rows) vs. our approach (bottom) for the wrestling motion. Top row shows the annotated pose-pairs in video, as input to direct 2D-to-3D lifting in second row [10] and third row [11]. Results in the bottom row are retrieved samples generated by our method, whose projections are close to the respective 2D pose-pairs in the top row. Quality and plausibility of the close interactions can be contrasted by focusing on the circled regions.

To the best of our knowledge, our work is the first to synthesize a large amount of 3D pose interaction data. We believe such data can be beneficial to data-driven solutions for the modeling and analysis of close human interactions. We demonstrate one such example in occluded joints inference for pose completion.

2 Related Work

Several topics relevant to our work, e.g., human motion capture, 2D-to-3D pose lifting, and character animation have been well-studied in computer graphics and computer vision. In this section, we focus on covering latest works that are most closely related.

2D-to-3D pose lifting. There have been many recent works on recovering *individual* 3D human poses from 2D data. Ramakrishna et al. [9] learn an over-complete set of basis poses from the CMU MoCap database and estimate 3D poses from 2D annotated joints as a sparse linear combination of the basis poses. Zhou et al. [11] propose a convex relaxation approach to solve for the sparse representation. Fan et al. [12] develop a pose locality constrained representation for 2D-to-3D pose lifting, which was also learned from the CMU MoCap database. Akhter and Black [10] learn a joint angle limit model from their new MoCap dataset that includes an extensive variety of stretching poses and use the learned model to constraint 2D-to-3D lifting of single poses. The quality of the lifted 3D poses can be enhanced by physics-based models. e.g., [13], [14]. More recently, Bogo et al. [15] estimate 2D joints automatically from single images and fit the 2D joints with a 3D statistical human model [16] to obtain 3D poses and body shapes.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2018.2832097, IEEE Transactions on Visualization and Computer Graphics

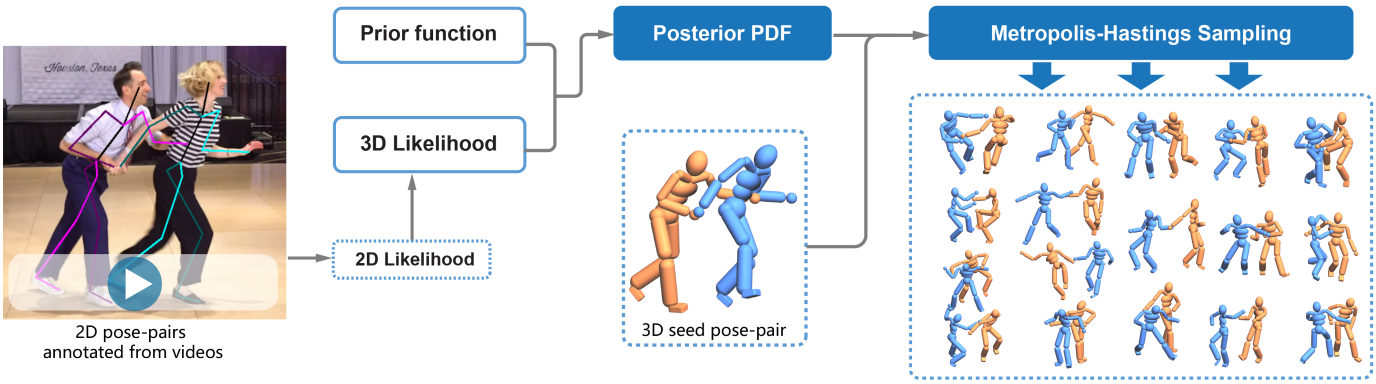


Figure 3. Overview of our 3D pose-pair generation via lifting-by-sampling.

To the best of our knowledge, there is no existing work yet on 2D-to-3D lifting, which is capable of recovering close interactions.

Sampling-based pose tracking. Sampling is also part of the solution pipeline for several methods for human pose tracking, where Bayesian distributions are also modeled from 2D observations. Deutscher et al. [17] propose annealed particle filtering for stochastic human tracking in high dimensional configuration spaces. Sminchisescu et al. [18] present a sample-and-refine searching strategy guided by rescaled cost-function covariances for 3D human body tracking from monocular video sequences. In a follow-up, [19] uses a discriminative density propagation model for human MoCap from 2D silhouettes. The fundamental difference between these methods and our lifting by sampling lies in whether the modeled distribution is *local* or *global*.

In particle filtering, the focus is to maintain a set of weighted particles (samples) to simulate the propagation of posterior distributions of a 3D human pose along the time dimension. At each time step, the particles are re-sampled after taking a new visual observation into account. As new observations come in, memory of old observations fades out, conforming to the local nature of the tracking problem. Generally, sampling-based human pose tracking aims to temporally model local distributions of single pose which respect the 2D observations and preserve temporal coherence. On the contrary, in our "lifting by sampling", we model one *global* posterior distribution given a type of close interacting 3D human pose-pairs. The goal of the sampling is to convert the dataset as a *whole* from 2D to 3D, i.e. lifting, instead of generating candidates that fit to an individual 2D observation.

Human interaction capture. Occlusions caused by close interactions have been a major challenge for human MoCap or pose estimation systems. Compared to the amount of literature on reconstructing or tracking single human motions, e.g., [7], [20], [21], [22], there have been considerably less works on capturing and estimating human interactions.

Liu et al. [23] propose a method to track multiple interacting characters, e.g., moving close to each other and hugging, with a setup of multiple cameras. In this work, a template of each person is tracked individually after performing a segmentation of their silhouettes. In a simpler acquisition setup, Ye et al. [24] employ three hand-held Microsoft Kinect Sensors to track light human interactions, again with the help of pre-captured human templates. Wang et al. [25] optimize a composite motion controller of a hand, interacting with a rigid object, to obtain physically realistic hand manipulation data from multi-view video. Using multiple optical sensors alleviates the occlusion problem to some extent, but would often reach its limit when human interactions are close, e.g., during wrestling. Inertial or magnetic systems may be free of data occlusion [26], [27], [28], but usually suffer from problems such as drifting or pose distortion. Utilizing deep learning techniques, recent works [29], [30] are able to detect non-occluded 2D joints of multiple persons from a single image, yet occluded joints remain a challenge.

Animating close human interactions. Animating moving characters with close interactions is a challenging problem. Past attempts include synthesizing multi-character motions from pre-captured single-character MoCap data [31], [32], combining physical simulation with real-time single-actor MoCap sequences to generate interacting motions with a virtual character [33], constructing coupled motion transition graphs and interaction models [34], and tackling the synthesis problem via motion patch tiling [35]. Earlier works by Ho et al. [36], [37] introduce topology coordinates to represent tangled limbs in their synthesis of character motions with close contact. In a follow-up [38], they employ Laplacian coordinates to adapt close interacting motions to skeletal configurations at varying scales. In the realm of interactive animation generation, users have been involved, e.g., to manipulate multi-character motions in both spatial and temporal domains [39], or to provide high-level descriptions and select preferred motions from candidates generated and ranked by an automated system [40]. In a recent work, Hyun et al. [41] predefine motion grammars with formal language, and synthesize animation of multi-person interactions, e.g., basketball playing, by a multi-level MCMC sampling approach. Most of the above methods rely on pre-captured motion data and are not well suited when the animated motions are difficult to capture to start with; this is the case with moving characters that are closely interacting.

3 Lifting by Sampling

Our goal is to generate a large and diverse set of closely interacting 3D human pose-pairs, for a given motion category, based on annotated video representing motions in that category. We treat this 2D-to-3D lifting task as a Markov chain Monte Carlo sampling problem, which can be solved using the well-established

Metropolis-Hastings algorithm; see Figure 3. The ensuing challenge is how to define and model the unknown probability density function (PDF) of *3D pose-pairs*. With the PDF, our Metropolis-Hastings algorithm starts from a single 3D seed pose-pair, traverse the space of plausible pose-pairs, and sample valid 3D pose-pairs with respect to the estimated posterior PDF.

Taking the wrestling motion as an example, where two characters are often closely interacting, we annotate each character over all available video frames with a 2D skeleton parameterized by 17 joints; see Figure 4(a). The premise is that we are able to evaluate the likelihood of the 2D pose-pairs for wrestling with this type of video annotations. To this end, we first estimate the camera view distribution from the annotated 2D data and based on this distribution, we lift the 2D likelihood function, via Monte Carlo integration, to a 3D likelihood function for 3D pose-pairs.

Specifically, we consider 3D pose-pair as a parameter set θ, which contains two root rotations, two root translations, and rotations of 32 non-root joints around their parent joints. Given a set D that contains 2D video annotations, a posterior probability density function $f(\theta|D)$ for 3D pose-pairs is constructed to be proportional to the product of a prior function and a likelihood function of θ:

$$f(\theta|D) \propto f(\theta) \cdot \mathcal{L}(\theta). \tag{1}$$

Here the prior function $f(\theta)$ is defined with the consideration of θ's physical plausibility, and the likelihood function $\mathcal{L}(\theta) = f(D|\theta)$ takes the set D as input and outputs θ's likelihood value.

4 Modeling Posterior PDF

In this section, we first define the prior function $f(\theta)$, the *physical prior* for pose-pairs, and then describe how to estimate the camera view distribution, and based on it how to estimate the 3D likelihood function $\mathcal{L}(\theta)$ from 2D annotations via lifting.

4.1 Physical pose prior

The *physical prior* in (1) is defined as:

$$f(\theta) \;=\; \begin{cases} 0, & \textit{if } isvalid(\theta) = 0, \\ \exp\left(-\frac{\|\theta-\bar{\theta}\|^2+\alpha\eta^2}{2\sigma_p^2}\right), & \textit{otherwise}, \end{cases} \tag{2}$$

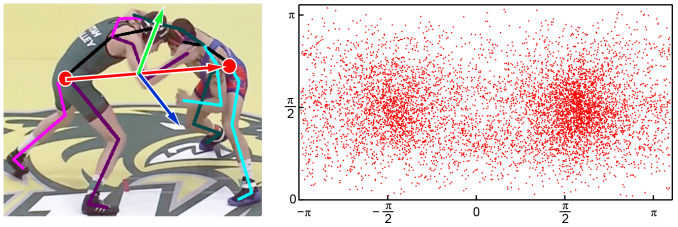
where $\bar{\theta}$ is the closest physically valid pose-pair to θ. We estimate $\bar{\theta}$ by applying the joint angle limits model [10] to the two single poses of a pair separately. The η denotes the sum of the penetration depths, detected by Open Dynamics Engine (ODE, [42]) among bones of the rigged character models. The weight $\alpha = 10$ and the bandwidth $\sigma_p = 0.05$ are set by default. We consider $isvalid(\theta) = 0$, i.e., definitely physically invalid, when the number of invalid joint angles is greater than a prescribed value, 8 by default in our implementation.

4.2 Camera view distribution

Given 2D annotated input data, the distribution of camera views in a local coordinate system is estimated first. The local coordinate system is defined with respect to two human poses from a single frame of annotation. To define the local coordinate system, we first apply the technique proposed by Akhter and Black [10] to reconstruct 3D backbones, i.e., the line segments that connect the chest joint and the hip joint, of the two characters from the same 2D annotation. To be more specific, we use the method to estimate 3D extended-torso [10] for each single pose of each 2D annotation frame. The 3D backbones are then calculated from the 3D extended-torsos. As 3D backbones are much easier and more robust to estimate than full 3D poses, we define local coordinate system with respect to two 3D backbones. Denoting these two backbones as B_1 and B_2 with two root points r_1 and r_2, we build a local coordinate system $(\mathbf{i}, \mathbf{j}, \mathbf{k})$ (shown in red, green and blue arrows in Figure 4(a)) for each data frame as:

$$\begin{aligned} \mathbf{i} &= N(r_2 - r_1), \\ \mathbf{j} &= N(N(B_1) + N(B_2)), \\ \mathbf{k} &= \mathbf{i} \times \mathbf{j}, \end{aligned} \tag{3}$$

where N denotes the operation of standard normalization, and $\mathbf{i}, \mathbf{j}, \mathbf{k}$ are column vectors.



(a) skeletal annotation (b) view distribution

Figure 4. Skeletal annotation on a wrestling video frame is shown in (a) with the defined local coordinate system. The plot (b) demonstrates the estimated camera views in 2D spherical coordinate $v = (\theta, \phi)$ for all 8732 wrestling video frames.

In the camera coordinate system, the camera view is the direction orthogonal to the plane of 2D annotation, i.e., $(0, 0, 1)^\top$. To convert the camera view into local coordinate system $(\mathbf{i}, \mathbf{j}, \mathbf{k})$, we solve a linear system $[\mathbf{i}, \mathbf{j}, \mathbf{k}](a, b, c)^\top = (0, 0, 1)^\top$, where $(a, b, c)^\top$ is the camera view in the local coordinate system. For more convenient computation, we convert $(a, b, c)^\top$ into 2D spherical coordinate $v = (\psi, \phi)$:

$$\psi = \arctan\left(\frac{c}{a}\right), \quad \phi = \arccos\left(\frac{b}{\sqrt{a^2+b^2+c^2}}\right), \tag{4}$$

where ψ and ϕ denote the angles between the view direction and the axis $\mathbf{i}$ and $\mathbf{j}$, respectively. Figure 4(b) presents the camera views estimated from annotated video frames of wrestling in spherical coordinates. We can observe that the views are denser around directions that are orthogonal to both $\mathbf{i}$ and $\mathbf{j}$ axes, which is consistent with what we would expect in real-life photography.

We can then naturally construct the probability density function for camera views via kernel density estimation (KDE):

$$f(v) = \frac{1}{n} \sum_{p_i \in D} g(v|v(p_i), \sigma_v^2 I), \tag{5}$$

where p_i is a 2D annotation in the input annotation set D, and $v(p_i)$ is the estimated camera view of p_i. The number of annotations in D is denoted by n, and $g(v|v(p_i), \sigma_v^2 I)$ is the Gaussian kernel with the mean $v(p_i)$ and the bandwidth σ_v (defaulted to 0.5 in radians). As we assume that the two variables of camera view are independent, the covariance matrix $\sigma_v^2 I$ is a diagonal matrix.

To sample $f(v)$, we first select a p_i from D uniformly at random and then draw the view sample from $g(v|v(p_i), \sigma_v^2 I)$.

4.3 Lifting via Interaction Coordinates (IC)

In order to estimate $\mathcal{L}(\theta)$, we introduce camera view v into the equation. Given the estimated camera view distribution $f(v)$ in (5), we obtain the 3D likelihood function $\mathcal{L}(\theta)$ by lifting a 2D likelihood function defined over 2D annotations via Monte Carlo integration over the camera view distribution:

$$\begin{aligned} \mathcal{L}(\theta) &= f(D|\theta) = \int f((D|\theta), v)dv \\ &= \int f(D|(\theta_v, v)) \cdot f(v)dv \\ &\approx \frac{1}{m}\sum_{i=1}^{m} f(D|(\theta_{v_i}, v_i)), \quad v_i \sim f(v), \end{aligned} \tag{6}$$

where v_i is a camera view sample, and θ_{v_i} denotes the 2D projection of 3D pose-pair θ under the view v_i. The decomposition by camera view makes the sampling problem more tractable. It allows us to sample pose-pairs in a local coordinate system without considering about the global coordinates of the pose-pairs.

To evaluate the likelihood function for (θ_{v_i}, v_i), that is:

$$\mathcal{L}(\theta_{v_i}, v_i) = f(D|(\theta_{v_i}, v_i)), \tag{7}$$

we need to measure the dissimilarity between two 2D pose-pairs in both single poses and the interactions between single poses.

Interaction coordinates. To this end, we propose a novel, integrated representation of a pair of 2D poses that are interacting closely, which we call *Interaction Coordinates* (IC). Specifically, IC offer a representation that encodes both poses and their interactions in an integrated manner. Given the Delaunay triangulation T of a 2D pose-pair θ, we define IC of the triangulated θ as an array of vectors on Delaunay edges, which yields:

$$I(\theta, T) = (..., u_{ij}(\theta) = J_i(\theta) - J_j(\theta), ...), (i,j) \in T,$$

where (i, j) is an edge of the Delaunay triangulation, and $J_i(\theta)$ denotes i-th joint of the pose-pair.

IC employ a graph-based global representation to jointly encode two single poses and the interactions between them. With IC, a human pose-pair is considered as a single unit in our sampling algorithm, where single poses and interaction are naturally balanced. Delaunay triangulation is a suitable choice for constructing this graph, since it maximizes the minimal angle in the triangulation so as to avoid parallel vectors in IC. Parallel vectors in triangulation give unbalanced high weights to the vertices connected by the vectors in measuring the difference between pose-pairs, thus they are not desirable here. Instead of using static corresponding relationships, our Delaunay edges are constructed dynamically for every pose-pair, so as to encode the interaction for every pose-pair.

With the definition of IC, we can now measure the dissimilarity between two 2D pose-pairs θ_1 and θ_2 as:

$$\begin{aligned} d_I(\theta_1, \theta_2) &= \frac{1}{W_1}\sum_{(i,j)\in T(\theta_1)} \left\| \frac{u_{ij}(\theta_1) - u_{ij}(\theta_2)}{max(\beta, ||u_{ij}(\theta_1)||)} \right\| \\ &+ \frac{1}{W_2}\sum_{(i,j)\in T(\theta_2)} \left\| \frac{u_{ij}(\theta_1) - u_{ij}(\theta_2)}{max(\beta, ||u_{ij}(\theta_2)||)} \right\|, \end{aligned} \tag{8}$$

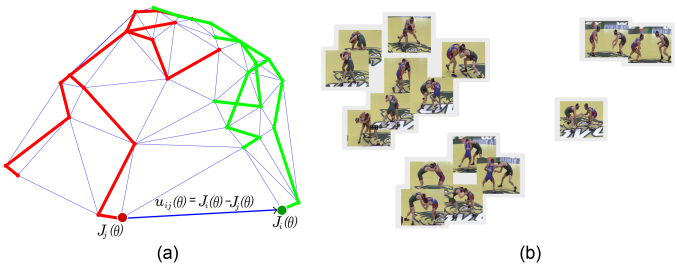


Figure 5. With IC illustrated in (a), we show the 2D embedding (b) using the dissimilarity distance defined in (8) over a small set of photos extracted from videos from which we annotated poses.

where the weight $1/max(\beta, ||u_{ij}(\theta)||)$ balances the impact of long and short Delaunay triangulation edges, with a parameter $\beta = 0.05$, by default. The summation of the weights is denoted by W. In the equation, we compute and compare the ICs twice, with two sets of Delaunay triangulation edges for the two 2D pose-pairs, since these two sets of edges are usually different.

Estimating likelihood using IC. Given the dissimilarity function d_I in (8) , the 2D likelihood function $\mathcal{L}(\theta_{v_i}, v_i)$ is estimated by finding a video annotation that is the closest one to θ_{v_i} :

$$\mathcal{L}(\theta_{v_i}, v_i) \propto \max_{p_j \in D} \exp\left(-\frac{d_I(p_j, M\theta_{v_i})}{2\sigma_l^2}\right), \tag{9}$$

where we set $\sigma_l = 0.01$ by default. In practice, it is expensive to consider all $p_j \in D$ for each pair of (θ_{v_i}, v_i). We thus only consider a subset of D that possess similar views to v_i.

To make the dissimilarity measure between two pose-pairs invariant to the global rotation and translation, we multiply θ_{v_i} with a least-square rigid transformation matrix M that aligns θ_{v_i} to p_j, where M is obtained via SVD [43].

To show the advantage of IC based dissimilarity measure, we demonstrate the 2D embedding of a set of 2D pose-pairs using $d_I(\cdot, \cdot)$ as distance metric; see Figure 5(b). The embedding is obtained with Multi-Dimensional Scaling (MDS).

5 Metropolis-Hastings Sampling

To draw 3D pose-pairs from the estimated posterior probability density function $f(\theta|D)$, Markov chain Monte Carlo (MCMC) sampling is applied using the Metropolis-Hastings algorithm [44]. To get a 3D seed pose-pair, we casually select a 2D annotation and convert the annotation to a 3D pose-pair by using single pose lifting technique [10]. The directly lifted 3D pose-pair is not of high quality. To allow the sampling algorithm to start from a location with high density value and thus reduce the number of low-quality samples, we manually edit a lifted example to obtain a high-quality 3D pose-pair as the seed. The Metropolis-Hastings sampling starts from a 3D seed pose-pair, randomly walks in the parameter space, generates a pose-pair θ' from the proposal distribution $Q(\theta'|\theta_i)$. Here θ' is accepted as a new sampled pose-pair θ_{i+1}, if

$$A(\theta'|\theta_i) = \frac{f(\theta'|D) * Q(\theta_i|\theta')}{f(\theta_i|D) * Q(\theta'|\theta_i)} \geq 1. \tag{10}$$

Otherwise, $A(\theta'|\theta_i)$ serves as a probability variable to accept θ'. With K seed pose-pairs, we may proceed with K Metropolis-Hastings sampling processes simultaneously. When the sampling

number increases greatly, the generated 3D new pose-pairs spread well and tend to cover the closely interacting poses observed from input video clips efficiently with novelty; see Figure 2.

5.1 Truncating the density function

Given a seed 3D pose-pair θ_0, we denote its density as f_0 and truncate the target density function using f_0 as reference:

$$\bar{f}(\theta|D) = \begin{cases} 0, & \text{if } f(\theta|D) < f_0/c, \\ cf_0, & \text{if } f(\theta|D) > cf_0, \\ f(\theta|D), & \text{otherwise}, \end{cases} \quad (11)$$

where the constant $c = 100$ by default. Note that $\bar{f}(\theta|D)$ uses the density of seed as a reference to set a reasonable range for density values of samples. We sample $\bar{f}(\theta|D)$ instead of $f(\theta|D)$ to keep away from outliers of very low density, and avoid redundant samples of very high density as well.

5.2 Velocity-biased proposal function

Gaussian distribution with preset variances is widely used as proposal distribution for Metropolis-Hastings algorithm. However, we notice that Gaussian proposal distribution with preset variance is not performing well in approximating $f(\theta'|D)$, as it violates the fact that the target density function $f(\theta'|D)$ tends to be locally constant along the moving velocity of θ' and descends steeply in the directions orthogonal to moving velocity. To better approximate the target density function, we propose a velocity-biased Gaussian proposal distribution, where the parameters with larger projections on velocity vector are assigned with larger variances. With the velocity-biased Gaussian proposal distribution, we can better approximate the target density function $f(\theta'|D)$, thus the efficiency of Metropolis-Hastings algorithm is promoted.

Predicting the velocity of an individual 3D pose-pair is clearly an ill-posed problem. Fortunately, we can again resort to the 2D annotated video which offers motion velocity data. From such data, we can directly estimate velocity-biased variances and then "lift" to 3D. Specifically, the variance $\sigma^2(k)$ of a component $\theta'(k)$ of the proposed parameter set θ' in velocity-biased Gaussian proposal distribution $Q(\theta'|\theta_i)$ is estimated as:

$$\sigma^2(k) = \sigma_0^2(k) + \frac{1}{|D^*|} \sum_{p \in D^*} |\Delta^p \theta_i(k)|^2, \quad (12)$$

where $\sigma_0(k)$ is a small initial variance assigned to $\theta'(k)$, and $\Delta^p \theta_i(k)$ is the local differential of 2D projection of $\theta_i(k)$ around a 2D annotation p in its video sequence. Here p is a member of the set D^*. Each 2D annotation in D^* is selected as the most similar annotation to one projection of θ_i, measured in IC space.

5.3 User control via MCMC restart

Although MCMC sampling is able to converge to the target density function after a finite number of sampling steps, we notice that not all users are interested in obtaining a huge amount of human poses from one seed. It is desirable to allow users to browse back the samples, choosing one they like, and then restarting MCMC from that sample to explore more from a new direction as they wish. Thus, we provide a playback and restart function at UI.

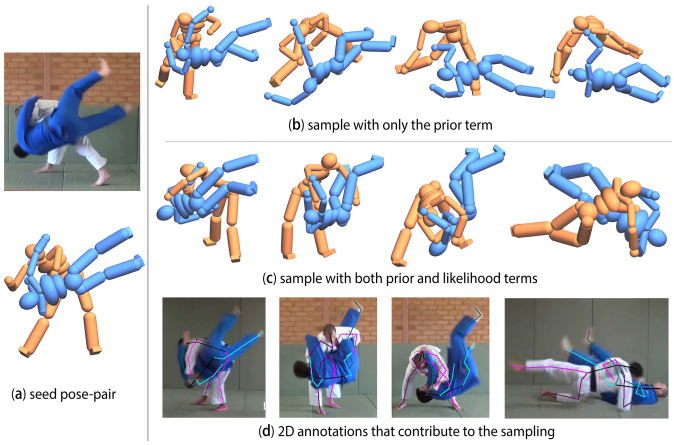


Figure 6. Ablation study of different terms in our probability density function that controls the MCMC sampling for Judo interactions. The 3D pose-pairs shown on the right, for each version of the PDF, represent the 200-th, 400-th, 600-th, and 800-th samples, respectively. To visually validate the results in (c), we also show examples of 2D annotations which contributed to the likelihood term in the bottom row (d).

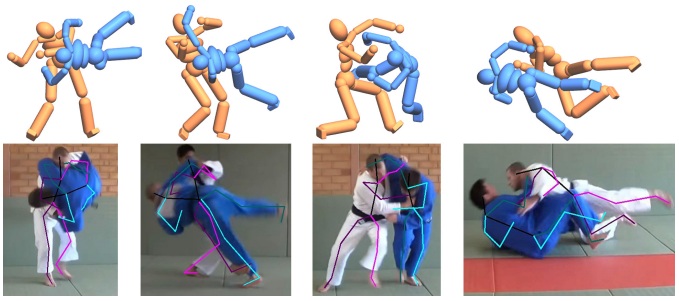


Figure 7. With an unbiased proposal function, the sampling algorithm tends to produce more diverse but less plausible pose-pairs. The first row shows the 200-th, 400-th, 600-th, and 800-th samples generated from the same seed shown in Figure 6(a). The second row shows examples of 2D annotations that contribute to the likelihood term.

6 Results, Evaluation, and Application

In this section, we show some results to validate and assess our MCMC sampling scheme for the generation of closely interacting 3D poses. First, we provide the results of an ablation study in Figure 6. We show visually the effects of the two terms in our probability density function (PDF) that controls the MCMC sampling. The motion category is Judo, for which we have collected and annotated 7,282 frames of video collected from on-line sources. The sampling starts with a single 3D pose-pair. We show 3D pose-pairs corresponding to the 200-th, 400-th, 600-th, and 800-th samples. As we can observe, the physical prior is able to bias the sampling towards producing physically valid individual poses. Adding the likelihood term improves the plausibility of the 3D pose-pairs as performing Judo motions. Note that when we sample with only the prior term, it is assumed that the algorithm does not rely on the annotated videos, thus the unbiased proposal function is used. When sampling with both the prior and likelihood terms, we always use the velocity-biased proposal function. Figure 6(d) shows 2D annotation examples that provided evidences to the likelihood of sampled pose-pairs. The pose-pairs in Figure 6(c) were rotated to match the views of the 2D annotations. With the same seed pose-pair, we show in Figure 7

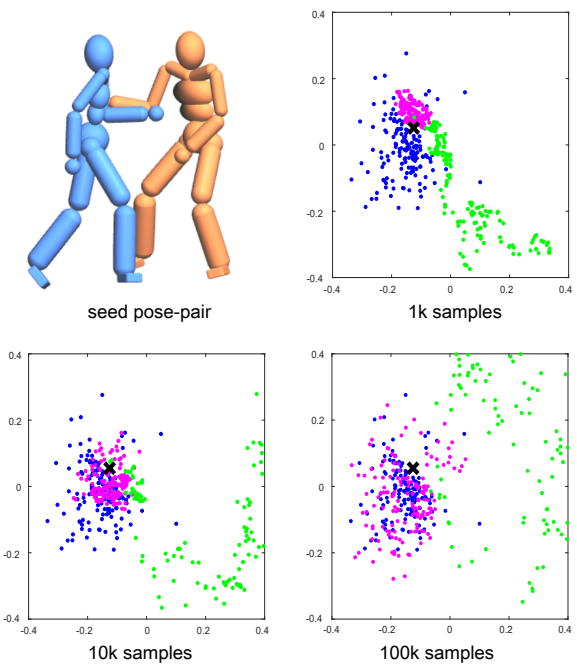


Figure 8. Sampling coverages via MDS embedding: our sampling scheme is shown in magenta dots over the space spanned by "ground truth" data given as blue dots; the green dots are pose-pairs generated by sampling with only a prior term. The number of samples increases from 1k through 10k to 100k. The seed is denoted by a black cross.

results of sampling with both the prior and likelihood terms using an unbiased proposal function. Compared to samples shown in Figure 6(c), the samples generated without considering the bias in the direction of velocity are more diverse. However, the samples are less constrained with respect to plausibility.

A key validation of our sampling scheme is how well and efficiently it can provide *coverage* of a space of 3D pose-pairs. For this purpose, we conduct an experiment using the Salsa dance dataset from the CMU MoCap database (http://mocap.cs.cmu.edu/). The Salsa dataset consists of a total of 15 sequences and about 31,000 frames of 3D pose-pairs performing Salsa dance. For our purpose, we removed pose-pairs that are clearly not performing the dance, e.g., initial poses for registration. Then we uniformly subsampled 6,000 frames to form the "ground-truth" (GT) data of 3D pose-pairs. The *space* of 3D pose-pairs which surround these GT data would be the target for our MCMC sampling to efficiently cover.

From this set of GT 3D pose-pairs, we randomly select and project along random views to produce 6,000 frames of 2D interacting pose-pairs, which form the "video" knowledge base for Salsa dance. We run our MCMC sampling with the PDF defined by this knowledge base, as described in Section 4. In Figure 8, through multi-dimensional scaling (MDS) visualization of the 3D pose-pairs, we visually demonstrate how well our samples (in magenta) are able to progressively "cover" the space of GT data (in blue). The embedding is obtained with a 3D version of the dissimilarity function $d_I(\theta_1, \theta_2)$. To demonstrate effectiveness of the likelihood term, we also plot the samples (in green) generated by sampling with the prior term only.

Let the set of GT 3D pose-pairs be $G = \{g_1, \ldots, g_m\}$ and let a

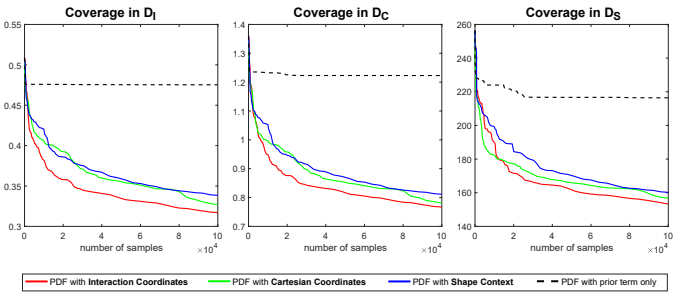


Figure 9. Plots of the coverage measurements in D_I, D_C, D_S, respectively. We compare the sampling with PDF defined over Interaction coordinates (red), Cartesian coordinates (green), shape context (blue), and PDF with only a prior term (black).

set of MCMC samples be $S = \{s_1, \ldots, s_k\}$, we can define the numerical measure of coverage as:

$$\mathcal{C}(G, S) = \frac{1}{m} \sum_{i=1}^{m} \mathcal{D}(g_i, s_i^*), \tag{13}$$

where $s_i^* \in S$ has the minimal dissimilarity $\mathcal{D}$ to g_i. We observe that the sampling coverage shown in Figure 8 is measured with respect to the space spanned by the GT data and not exclusively to the GT data itself.

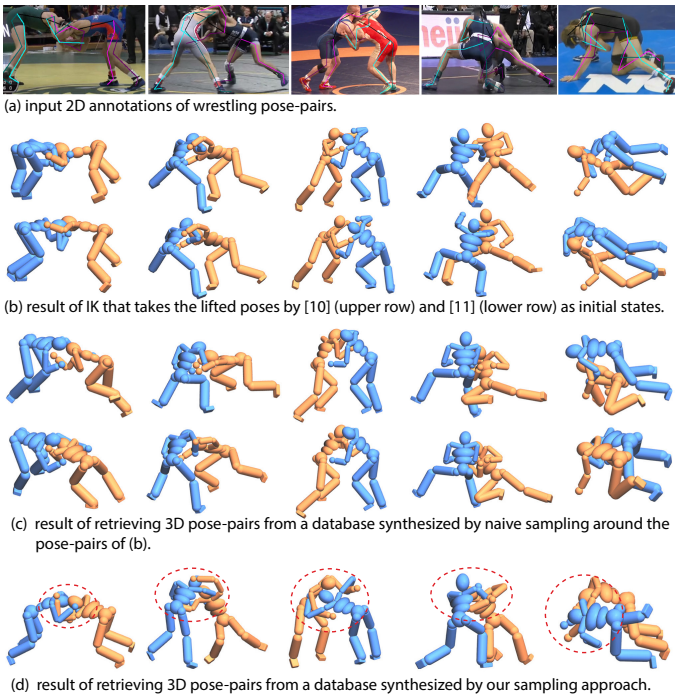
In Figure 9, we numerically evaluate the sampling coverage with three different dissimilarity measures. The first metric is denoted as D_I, which is a 3D version of the dissimilarity function defined for our interaction coordinates (8). To extend our IC definition from 2D to 3D, we naturally turn Delaunay triangulations in the plane to Delaunay tetrahedralizations in 3D space. Thus, the second metric corresponds to Euclidean distance in Cartesian coordinates of the 3D pose-pairs, denoted by D_C. The third one, denoted as D_S, is the matching error in 3D shape contexts [45], one of the best known and most widely applied shape descriptor.

We plot the three coverage measures as we increase the number of MCMC samples. The coverages are measured for the samplings with four different versions of PDF. The PDF defined over IC (red curve) is the default configuration for our method. Figure 9 shows that it performs better in covering the GT space than PDFs defined over Cartesian Coordinates (green) and shape contexts (blue), thanks to its ability in encoding both poses and their interactions in an integrated manner. As a baseline for our comparison, we also plot the coverage (black) obtained by sampling with only a prior term.

To visually demonstrate the in-between poses generated from a single seed by our sampling schema is diverse and plausible, we show an example in Figure 11, where three closest 2D annotations are provided to the right side of each sampled 3D pose-pair to demonstrate the validity of samples.

Comparisons to enhanced lifting. We provide comparisons to two baseline methods in Figure 10, where the same examples as shown in Figure 2 are used. The first method is lifting+IK, for which we take the 2D-to-3D single pose lifting results, as shown in Figure 2, as initial states of the IK system, and minimize an energy $E_{2D}+E_{isvalid}$ while the bone lengths and interpenetration resolution are strictly constrained. In the energy function, E_{2D} measures $L2$ distance from the 2D projection of 3D pose-pair to input 2D annotation. $E_{isvalid}$ measures $L2$ distance from the 3D

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2018.2832097, IEEE Transactions on Visualization and Computer Graphics



(a) input 2D annotations of wrestling pose-pairs.

(b) result of IK that takes the lifted poses by [10] (upper row) and [11] (lower row) as initial states.

(c) result of retrieving 3D pose-pairs from a database synthesized by naive sampling around the pose-pairs of (b).

(d) result of retrieving 3D pose-pairs from a database synthesized by our sampling approach.

Figure 10. Comparisons to enhanced lifting (b) and a naive sampling (c) around the 3D pose-pairs generated by enhanced lifting. For a clear assessment of the comparisons, we also show 2D annotations (a) and our results (d) that are the same as shown in Figure 2.

pose-pair to closest valid poses suggested by the joint angle limits model. The optimization is solved by the method of Lagrange multipliers. The result is provide in Figure 10(b).

For each of the 3D pose-pairs shown in Figure 10(b), we randomly sample 1000 new pose-pairs within a local neighborhood around it to construct a database. During the sampling, we strictly constrain the bone lengths and interpenetration resolution. A bias is given to pose-pairs that are closer to physically valid poses in terms of the joint angle limits model. We retrieve for a closest pose-pair to the 2D annotation in Figure 10(a) from the database, and provide them in Figure 10(c). For comparison, the results of retrieving from database generated by our own method are shown in the bottom row, i.e., Figure 10(d). Overall, our method is more plausible in terms of interactions between two poses.

The way IK works in the enhanced lifting approach is essentially to apply a new prior on single human poses. One of the reasons why our method achieves better results is that it considers both single human pose priors and a data-driven term defined over a set of 2D pose-pair annotations. Moreover, while the enhanced lifting approach only works in the space of single human poses, our approach also models the interactions between two human poses with the proposed interaction coordinates (IC).

Occluded joint inference. The 3D human pose-pairs generated via lifting-by-sampling in this work can effectively facilitate the recovery of closely interacting motions. The key challenge of reconstructing such motions arises from significant occlusions. In Figure 12, we provide an example to show that the 3D pose-pairs we obtain can lead to better joint inference and completion of highly occluded poses captured by inexpensive depth sensors.

Specifically, we employ a simple retrieval-based solution for the pose completion task. Given an incomplete pose-pair $\mathcal{P}$, we search for a small subset (e.g., 100, as the default in our experiment) of pose-pairs, which are close to $\mathcal{P}$ in metric D_I, from the sample database we generated. We then estimate a median pose-pair $\mathcal{M}$ of the subset. Finally, we deform $\mathcal{M}$ to $\mathcal{P}$ via Gaussian-weighted linear blending to obtain the completion result. Noting that, without any smoothing operation applied in the temporal domain, the motions that we completed demonstrate sufficient temporal coherence for the visual perception. Please refer to the supplementary demo video for a detailed visual demonstration.

Implementation details. We implemented the PDF and sampling algorithm with C++. The sampling speed is around 10 fps on an Intel i7 4-core 3.4GHz CPU. During the sampling, the local motion of each non-root joint has three degrees of freedom, including two rotations around its parent joint and one translation along its bone. Ten volunteers participated in the video data annotation. With a semi-automatic annotation tool that involves simple joint tracking and interpolation, the average time one volunteer spent on annotating one single video frame was around 20 seconds.

7 Summary, Limitations, and Future Work

We have presented a method to generate closely interacting 3D pose-pairs, which offers a means to augment few, or even a single, seed pose-pair(s) with a large number of synthesized pose-pairs sampled from a PDF model. The main challenge was to model the PDF. Our idea is to first estimate the distribution of camera views from annotated video frames, then with the known view distribution we can lift the density estimated from 2D data to the density for 3D pose-pairs with Monte Carlo estimation. The PDF of 3D pose-pairs is modeled as a combination of two density functions that consist of a physical validity term and an interaction plausibility term.

The close interaction between the two bodies is challenging mainly due to inter-pose as well as self-occlusions, but the intertwining arms or body parts give rise to the a constrained search space. We developed interaction coordinates to encode the interactions between two 2D poses. This representation, while improving upon the classical Cartesian coordinates representation, is still quite elementary. Also, there is no strong reason to believe that Delaunay triangulations are the most suitable structure to connect the relevant joints. Other choices including Knn graphs, Gabriel graphs, and minimum weight triangulations could also be experimented with. Generally, it remains to be further investigated whether there are stronger representations of the closely interacting body parts that may be more descriptive and effective.

The key idea of *lifting by sampling* bypasses the challenges in directly lifting a 2D pose to 3D. The inherent ambiguity in the projected image of a pose-pair is significantly higher than the notorious ambiguity in a single body. By applying a Markov Chain Monte Carlo (MCMC) method, we alleviate the problem, taking advantage of the rather dense space of annotated video frames. We believe that our technique can be effective also in similar problems including modeling or analyzing two-hand postures. In this scenario, the amount of self-occlusion can also be extremely high, preventing a reliable reconstruction from a single view.

The main limitation of our method is that the plausibility is data-driven, and the generated pose-pairs are sampled around those

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2018.2832097, IEEE Transactions on Visualization and Computer Graphics

Figure 11. From a single seed, our sampling schema produces diverse pose-pairs of wrestling. Three closest 2D annotations are provided to the right side of each sampled 3D pose-pair to demonstrate its validity.

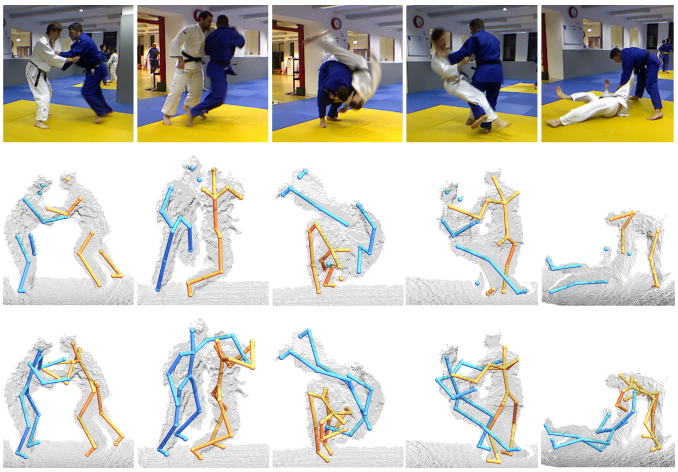


Figure 12. Pose completion results for highly occluded human pose-pairs annotated from depth images, which were captured using Kinect (the middle row). The bottom row presents the 3D pose-pairs completed by a simple retrieval-based solution using our samples.

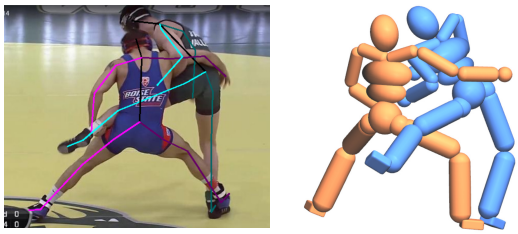


Figure 13. Our current pose annotation (left) does not include front-back or depth information, which may lead to ambiguities (right).

observed and annotated video frames only. The challenge then is to generate a richer set, more diverse than the one annotated. One can consider applying simple transformations and local perturbations over the 3D pose-pairs themselves, however, this must be executed with care to ensure their plausibility. We plan to enhance the plausibility of pose-pairs using inverse-kinematics techniques applied to each body separately and to the two bodies, possibly forming a connected skeleton with virtual links connecting them. It is also possible to edit the poses while preserving the spatial relationship between the body parts using [38].

Another technical limitation arises from a lack of front-back information associated with the interacting poses from 2D annotations; see Figure 13. The ensuing ambiguities may cause the generation of implausible 3D pose-pairs under the current PDF. As well, we currently do not consider grounding of the characters when performing the motions or factor that into the PDF. At last, we do not yet have an efficient organization of the large set of generated 3D pose-pairs to facilitate pose-pair retrievals that are necessary for the *lifting-by-sampling* and *interpolating-by-sampling* tasks.

We believe that similar data augmentation problems will receive more attention, especially in the context of deep learning, where larger amount of data is required for training. The need for such techniques is more acute for 3D data, like the one we are dealing with, which is hard to acquire and annotate or contains occlusions and inherent ambiguities. Finally, it is also a compelling problem to generalize interaction coordinates to encode interactions among multiple people and explore the potential of our sampling approach to generate *multi-interaction motions*.

Acknowledgment

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions. This work was supported in part by NSFC (61522213, 61761146002, 6171101466), 973 Program (2015CB352501), Guangdong Science and Technology Program (2015A030312015), Shenzhen Innovation Program (KQJSCX20170727101233642, JCYJ20151015151249564), ISF-NSFC Joint Research Program (2472/17) and NSERC (611370).

References

[1] A. Lamouret and M. van de Panne, "Motion synthesis by example," in *Eurographics Workshop on Computer Animation and Simulation*, 1996, pp. 199–212.

[2] J. Shotton, A. W. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.*, 2011, pp. 1297–1304.

[3] Y. Lee, S. Kim, and J. Lee, "Data-driven biped control," *ACM Trans. on Graph (SIGGRAPH)*, vol. 29, no. 4, p. 129, 2010.

[4] V. G. Kim, S. Chaudhuri, L. Guibas, and T. Funkhouser, "Shape2Pose: Human-centric shape analysis," *ACM Trans. on Graph (SIGGRAPH)*, vol. 33, no. 4, pp. 120:1–12, 2014.

[5] M. Savva, A. X. Chang, P. Hanrahan, M. Fisher, and M. Nießner, "Scenegrok: Inferring action maps in 3d environments," *ACM Trans. on Graph*, vol. 33, no. 6, pp. 212:1–10, 2014.

[6] R. Hu, C. Zhu, O. van Kaick, L. Liu, A. Shamir, and H. Zhang, "Interaction context (icon): Towards a geometric functionality descriptor," *ACM Trans. on Graph (SIGGRAPH)*, vol. 34, no. 4, p. Article 83, 2015.

[7] X. Wei, P. Zhang, and J. Chai, "Accurate realtime full-body motion capture using a single depth camera," *ACM Trans. on Graph (SIGGRAPH)*, vol. 31, no. 6, p. 188, 2012.

[8] C. J. Taylor, "Reconstruction of articulated objects from point correspondences in a single uncalibrated image," in *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.*, vol. 1. IEEE, 2000, pp. 677–684.

[9] V. Ramakrishna, T. Kanade, and Y. Sheikh, "Reconstructing 3d human pose from 2d image landmarks," in *Proc. Euro. Conf. on Comp. Vis.* Springer, 2012, pp. 573–586.

[10] I. Akhter and M. J. Black, "Pose-conditioned joint angle limits for 3d human pose reconstruction," in *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.*, 2015, pp. 1446–1455.

[11] X. Zhou, S. Leonardos, X. Hu, and K. Daniilidis, "3d shape estimation from 2d landmarks: A convex relaxation approach," in *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.*, 2015, pp. 4447–4455.

[12] X. Fan, K. Zheng, Y. Zhou, and S. Wang, "Pose locality constrained representation for 3d human pose reconstruction," in *Proc. Euro. Conf. on Comp. Vis.* Springer, 2014, pp. 174–188.

[13] X. Wei and J. Chai, "Videomocap: modeling physically realistic human motion from monocular video sequences," *ACM Trans. on Graph (SIGGRAPH)*, vol. 29, no. 4, p. 42, 2010.

[14] M. Vondrak, L. Sigal, J. Hodgins, and O. Jenkins, "Video-based 3d motion capture through biped control," *ACM Trans. on Graph (SIGGRAPH)*, vol. 31, no. 4, p. 27, 2012.

[15] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it smpl: Automatic estimation of 3d human pose and shape from a single image," in *Proc. Euro. Conf. on Comp. Vis.*, 2016, pp. 561–578.

[16] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: a skinned multi-person linear model," *ACM Trans. on Graph (SIGGRAPH Asia)*, vol. 34, no. 6, p. 248, 2015.

[17] J. Deutscher, A. Blake, and I. Reid, "Articulated body motion capture by annealed particle filtering," in *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.*, vol. 2. IEEE, 2000, pp. 126–133.

[18] C. Sminchisescu and B. Triggs, "Covariance scaled sampling for monocular 3d body tracking," in *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.*, vol. 1. IEEE, 2001.

[19] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas, "Discriminative density propagation for 3d human motion estimation," in *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.*, vol. 1. IEEE, 2005, pp. 390–397.

[20] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore, "Real-time human pose recognition in parts from single depth images," *Communications of the ACM*, vol. 56, no. 1, pp. 116–124, 2013.

[21] A. Baak, M. Müller, G. Bharaj, H.-P. Seidel, and C. Theobalt, "A data-driven approach for real-time full body pose reconstruction from a depth camera," in *Consumer Depth Cameras for Computer Vision*. Springer, 2013, pp. 71–98.

[22] P. Zhang, K. Siu, J. Zhang, C. K. Liu, and J. Chai, "Leveraging depth cameras and wearable pressure sensors for full-body kinematics and dynamics capture," *ACM Trans. on Graph (SIGGRAPH Asia)*, vol. 33, no. 6, p. 221, 2014.

[23] Y. Liu, C. Stoll, J. Gall, H.-P. Seidel, and C. Theobalt, "Markerless motion capture of interacting characters using multi-view image segmentation," in *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.*, 2011, pp. 1249–1256.

[24] G. Ye, Y. Liu, N. Hasler, X. Ji, Q. Dai, and C. Theobalt, "Performance capture of interacting characters with handheld kinects," in *Proc. Euro. Conf. on Comp. Vis.*, 2012, pp. 828–841.

[25] Y. Wang, J. Min, J. Zhang, Y. Liu, F. Xu, Q. Dai, and J. Chai, "Video-based hand manipulation capture through composite motion control," *ACM Trans. on Graph (SIGGRAPH)*, vol. 32, no. 4, p. 43, 2013.

[26] J. Morris, "Accelerometry a technique for the measurement of human body movements," *Journal of biomechanics*, vol. 6, no. 6, pp. 729–736, 1973.

[27] F. H. Raab, E. B. Blood, T. O. Steiner, and H. R. Jones, "Magnetic position and orientation tracking system," *IEEE Transactions on Aerospace and Electronic systems*, no. 5, pp. 709–718, 1979.

[28] D. Roetenberg, H. Luinge, and P. Slycke, "Xsens mvn: full 6dof human motion tracking using miniature inertial sensors," *Xsens Motion Technologies BV, Tech. Rep*, 2009.

[29] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler, and B. Schiele, "Deepcut: Joint subset partition and labeling for multi person pose estimation," in *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.*, June 2016.

[30] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," *arXiv preprint arXiv:1611.08050*, 2016.

[31] C. K. Liu, A. Hertzmann, and Z. Popović, "Composition of complex optimal multi-character motions," in *Proc. ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. Eurographics Association, 2006, pp. 215–222.

[32] H. P. Shum, T. Komura, M. Shiraishi, and S. Yamazaki, "Interaction patches for multi-character animation," in *ACM Trans. on Graph (SIGGRAPH Asia)*, vol. 27, no. 5. ACM, 2008, p. 114.

[33] N. Nguyen, N. Wheatland, D. Brown, B. Parise, C. K. Liu, and V. Zordan, "Performance capture with physical interaction," in *Proc. ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. Eurographics Association, 2010, pp. 189–195.

[34] T. Kwon, Y.-S. Cho, S. I. Park, and S. Y. Shin, "Two-character motion analysis and synthesis," *IEEE Trans. Vis. & Comp. Graphics*, vol. 14, no. 3, pp. 707–720, 2008.

[35] K. Hyun, M. Kim, Y. Hwang, and J. Lee, "Tiling motion patches," *IEEE Trans. Vis. & Comp. Graphics*, vol. 19, no. 11, pp. 1923–1934, 2013.

[36] E. S. Ho and T. Komura, "Character motion synthesis by topology coordinates," in *Computer Graphics Forum (Eurographics)*, vol. 28, no. 2, 2009, pp. 299–308.

[37] E. S. Ho and T. Komura, "Indexing and retrieving motions of characters in close contact," *IEEE Trans. Vis. & Comp. Graphics*, vol. 15, no. 3, pp. 481–492, 2009.

[38] E. S. Ho, T. Komura, and C.-L. Tai, "Spatial relationship preserving character motion adaptation," *ACM Trans. on Graph (SIGGRAPH)*, vol. 29, no. 4, p. 33, 2010.

[39] M. Kim, K. Hyun, J. Kim, and J. Lee, "Synchronized multi-character motion editing," in *ACM Trans. on Graph (SIGGRAPH)*, vol. 28, no. 3. ACM, 2009, p. 79.

[40] J. Won, K. Lee, C. O'Sullivan, J. K. Hodgins, and J. Lee, "Generating and ranking diverse multi-character interactions," *ACM Trans. on Graph (SIGGRAPH Asia)*, vol. 33, no. 6, p. 219, 2014.

[41] K. Hyun, K. Lee, and J. Lee, "Motion grammars for character animation," in *Computer Graphics Forum*, vol. 35, no. 2. Wiley Online Library, 2016, pp. 103–113.

[42] R. L. Smith, "Open dynamics engine," http://www.ode.org/, 2001–2004.

[43] K. S. Arun, T. S. Huang, and S. D. Blostein, "Least-squares fitting of two 3-d point sets," *IEEE Trans. Pat. Ana. & Mach. Int.*, no. 5, pp. 698–700, 1987.

[44] W. K. Hastings, "Monte carlo sampling methods using markov chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, 1970.

[45] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Trans. Pat. Ana. & Mach. Int.*, vol. 24, no. 4, pp. 509–522, 2002.

Kangxue Yin is currently a Ph.D. student with the school of Computing Science, Simon Fraser University. Before that, he received the bachelor degree in software engineering from Chang'an University in 2012, and worked as research assistant at Shenzhen Institute of Advanced Technology from 2012 to 2015. His research interest is in computer graphics.

Hui Huang is a Distinguished Professor of Shenzhen University, where she directs the Visual Computing Research Center in the College of Computer Science and Software Engineering. She received her Ph.D. in Applied Math from The University of British Columbia in 2008 and another Ph.D. in Computational Math from Wuhan University in 2006. Her research interests are in computer graphics and scientific computing, focusing on point-based modeling, geometric analysis, 3D acquisition and creation.

Edmond. S.L. Ho received the Ph.D. degree from the University of Edinburgh, Scotland, in 2011. He is currently a senior lecturer with the Department of Computer and Information Sciences, Northumbria University, United Kingdom. His current research interests include character animation, robotics, and human activity understanding.

Hao Wang is a postdoctoral researcher in the Visual Computing Research Center at Shenzhen University. He received his Ph.D. degree from Huazhong University of Science and Technology in 2016. His primary research lies in Computer Graphics, including mesh segmentation and ocean simulation.

Taku Komura is a Reader (Associate Professor) at the Institute of Perception, Action and Behaviour, School of Informatics, Edinburgh University. He is also a Royal Society Industry Fellow and a visiting professor at Xi'an Jiaotong University. He received his B.Sc., M.Sc. and D.Sc. in Information Science from the University of Tokyo. His research interests include character animation, computer graphics and interactive techniques.

Daniel Cohen-Or is a Professor in the Department of Computer Science at Tel Aviv University. He received a B.Sc. degree in both Mathematics and Computer Science (1985), a M.Sc. degree in Computer Science (1986) from Ben-Gurion University, and a Ph.D. from the Department of Computer Science (1991) at State University of New York at Stony Brook. His research interests are in Computer Graphics, Visual Computing and Geometric Modeling and including rendering and modeling techniques, shape analysis, shape creation and editing, 3D reconstruction, photo processing, compression and streaming techniques, visibility, point set representation, morphing and volume graphics.

Hao Zhang is a full professor in the School of Computing Science at Simon Fraser University (SFU), Canada, where he directs the graphics (GrUVi) lab. He obtained his Ph.D. from the Dynamic Graphics Project (DGP), University of Toronto, and M.Math. and B.Math degrees from the University of Waterloo, all in computer science. Richard's research is in computer graphics with a focus on geometry modeling, shape analysis, 3D content creation, and computational design and fabrication.